


Writing a (very niche) R package for reproducible data analysis from a public dataset



Anh N. Tran

Postdoc at Northwestern U.

@anh_n_tran 

trannhatanh89 

SatRday Chicago, April 27, 2019

PAX8

Search Fields »



TISSUE ATLAS

PRIMARY DATA

GENE/PROTEIN

ANTIBODIES
AND
VALIDATION

Dictionary



Tissue proteome



GENERAL INFORMATION

Gene name¹ PAX8
Gene description¹ Paired box 8
Protein class¹ Cancer-related genes
Disease related genes
Predicted intracellular proteins
Transcription factors
Predicted localization¹ Intracellular
Number of transcripts¹ 9

SHOW MORE

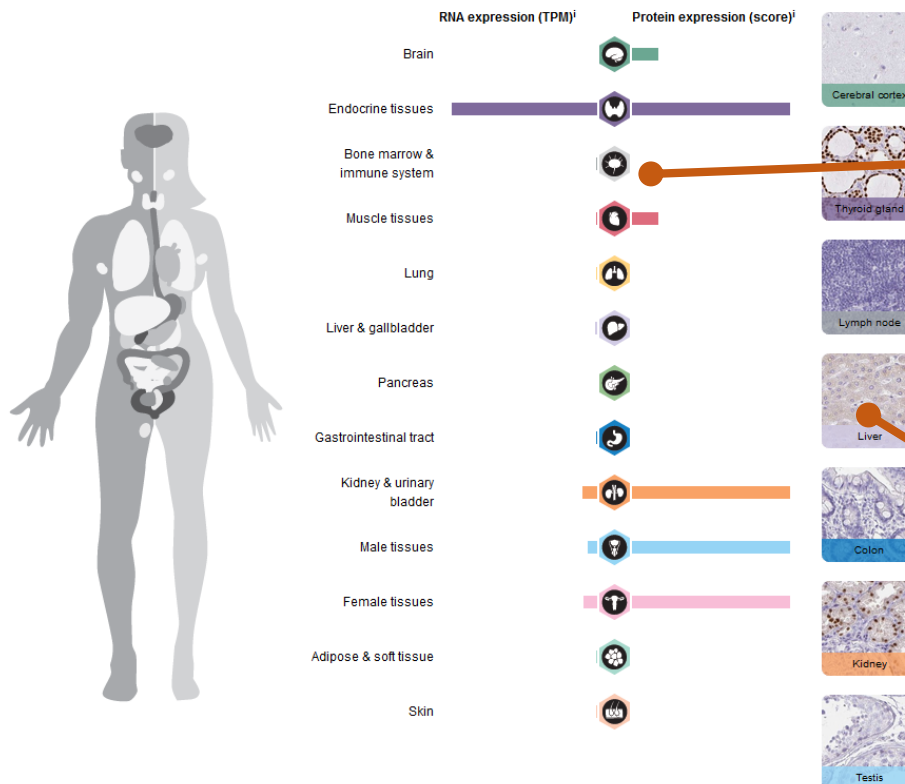
HUMAN PROTEIN ATLAS INFORMATION

RNA tissue category¹ HPA: Tissue enriched (thyroid gland)
GTEx: Group enriched (kidney, thyroid gland)
FANTOM5: Group enriched (kidney, thyroid gland)
Protein evidence¹ Evidence at protein level
Protein expression¹ Nuclear expression in thyroid glandular cells and renal tubuli cells.

IMMUNOHISTOCHEMISTRY DATA RELIABILITY

Data reliability description¹ Antibody staining consistent with RNA expression data.
Reliability score¹ Enhanced
Antibodies¹ CAB055097, CAB061888

SHOW MORE

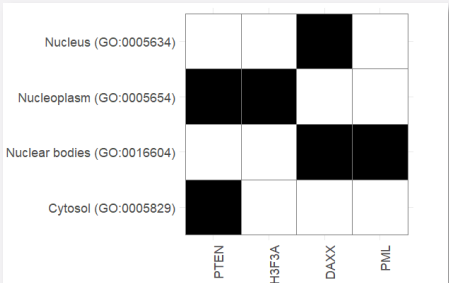
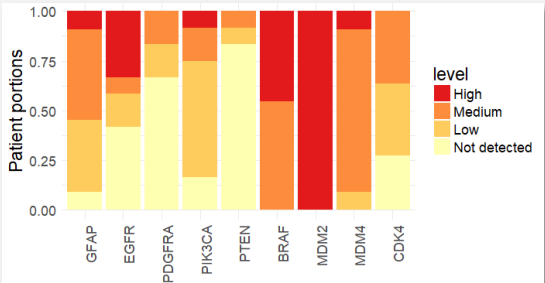
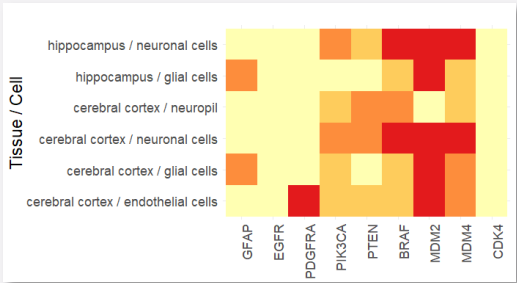
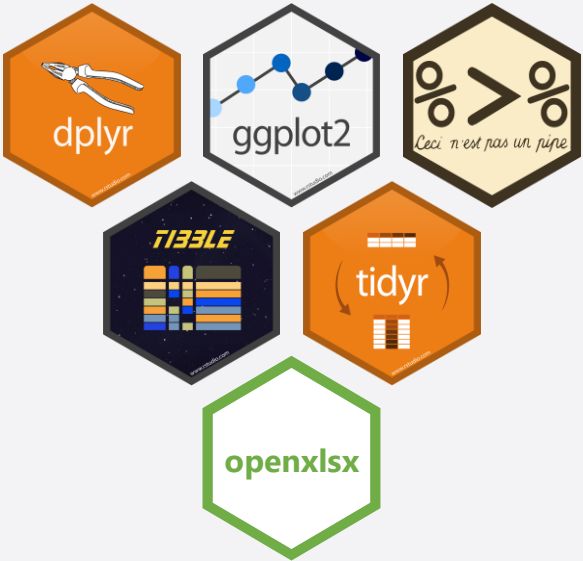
RNA AND PROTEIN EXPRESSION SUMMARY¹

Full of info, but untidy

Graphs are pretty,
but not publication
qualityHard to retrieve data
for reproducibility

Summarized data provided as .tsv

Gene	Gene name	Tissue	Cell type	Level	Reliability
ENSG000000000003	TSPAN6	adrenal gland	glandular cells	Not detected	Approved
ENSG000000000003	TSPAN6	appendix	glandular cells	Medium	Approved
ENSG000000000003	TSPAN6	appendix	lymphoid tissue	Not detected	Approved
ENSG000000000003	TSPAN6	bone marrow	hematopoietic cells	Not detected	Approved
ENSG000000000003	TSPAN6	breast	adipocytes	Not detected	Approved
ENSG000000000003	TSPAN6	breast	glandular cells	High	Approved
ENSG000000000003	TSPAN6	breast	myoepithelial cells	Not detected	Approved
ENSG000000000003	TSPAN6	bronchus	respiratory epithelial cells	High	Approved
ENSG000000000003	TSPAN6	caudate	glial cells	Not detected	Approved
ENSG000000000003	TSPAN6	caudate	neuronal cells	Not detected	Approved
ENSG000000000003	TSPAN6	cerebellum	cells in granular layer	Not detected	Approved
ENSG000000000003	TSPAN6	cerebellum	cells in molecular layer	Not detected	Approved
ENSG000000000003	TSPAN6	cerebellum	Purkinje cells	Not detected	Approved
ENSG000000000003	TSPAN6	cerebral cortex	endothelial cells	Not detected	Approved
ENSG000000000003	TSPAN6	cerebral cortex	glial cells	Not detected	Approved
ENSG000000000003	TSPAN6	cerebral cortex	neuronal cells	Medium	Approved
ENSG000000000003	TSPAN6	cerebral cortex	neuropil	Not detected	Approved
ENSG000000000003	TSPAN6	cervix, uterine	glandular cells	High	Approved
ENSG000000000003	TSPAN6	cervix, uterine	squamous epithelial cells	High	Approved
ENSG000000000003	TSPAN6	colon	endothelial cells	Not detected	Approved
ENSG000000000003	TSPAN6	colon	glandular cells	Medium	Approved
ENSG000000000003	TSPAN6	colon	peripheral nerve/ganglion	Not detected	Approved
ENSG000000000003	TSPAN6	duodenum	glandular cells	Low	Approved
ENSG000000000003	TSPAN6	endometrium 1	cells in endometrial stroma	Not detected	Approved
ENSG000000000003	TSPAN6	endometrium 1	glandular cells	High	Approved
ENSG000000000003	TSPAN6	endometrium 2	cells in endometrial stroma	Not detected	Approved

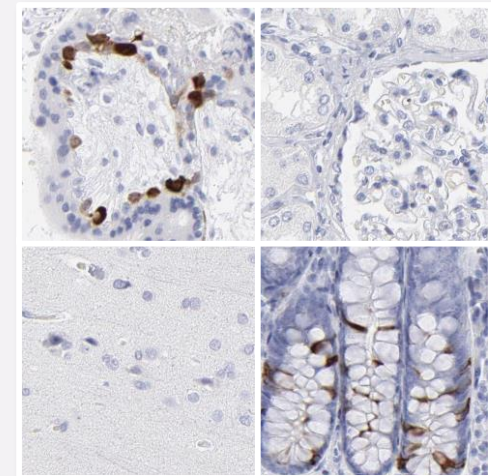


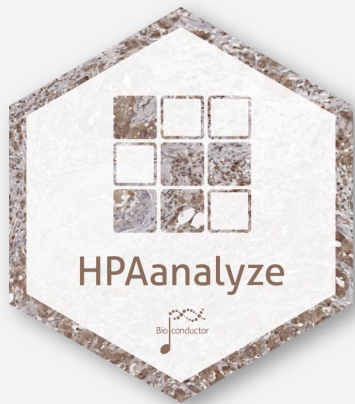
Full datasets provided as .xml(s)

```
<proteinAtlas xsi:schemaLocation="http://v18.proteinatlas.org/download/proteinatlas.xsd" schemaVersion="2.5">
  <entry version="18" url="http://v18.proteinatlas.org/ENSG00000134057">
    <name>CCNB1</name>
    <synonym>CCNB</synonym>
    <identifier id="ENSG00000134057" db="Ensembl" version="88.38">
      <xref id="P14635" db="Uniprot/SWISSPROT"/>
    </identifier>
    <proteinClasses>
      <proteinClass source="TCDB" id="Ja" parent_id="" name="Transporters"/>
      <proteinClass source="TCDB" id="Jb" parent_id="Ja" name="Transporter channels and pores"/>
      <proteinClass source="SPOCTOPUS" id="Mi" parent_id="" name="SPOCTOPUS predicted membrane proteins"/>
      <proteinClass source="HPA" id="Za" parent_id="" name="Predicted intracellular proteins"/>
      <proteinClass source="Plasma Proteome Database" id="Pp" parent_id="" name="Plasma proteins"/>
      <proteinClass source="" id="Ca" parent_id="" name="Cancer-related genes"/>
      <proteinClass source="Plasma Proteome Institute" id="Cb" parent_id="Ca" name="Candidate cancer biomarkers"/>
      <proteinClass source="UniProt" id="Ua" parent_id="" name="UniProt - Evidence at protein level"/>
      <proteinClass source="Kim et al 2014" id="Ea" parent_id="" name="Protein evidence (Kim et al 2014)"/>
      <proteinClass source="Ezkurdia et al 2014" id="Eb" parent_id="" name="Protein evidence (Ezkurdia et al 2014)"/>
    </proteinClasses>
    <proteinEvidence evidence="Evidence at protein level">
      <evidence source="HPA" evidence="Evidence at protein level"/>
      <evidence source="MS" evidence="Evidence at protein level"/>
      <evidence source="UniProt" evidence="Evidence at protein level"/>
    </proteinEvidence>
    <tissueExpression source="HPA" technology="IHC" assayType="tissue">
      <summary type="tissue">Cytoplasmic expression in proliferating cells.</summary>
      <verification type="reliability" description="Antibody staining mainly consistent with RNA expression data.">enhanced</verification>
      <image imageType="selected">
        <tissue>cerebral cortex</tissue>
        <imageUrl>
          http://v18.proteinatlas.org/images/3804/10580_B_8_5_rna_selected.jpg
        </imageUrl>
      </image>
    </tissueExpression>
  </entry>
</proteinAtlas>
```

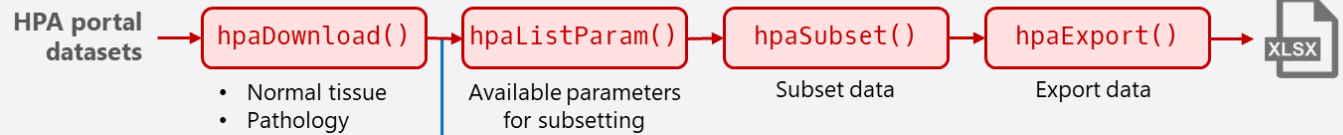


```
#> # A tibble: 327 x 18
#>   patientId age sex staining intensity quantity location imageUrl
#>   <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
#> 1 1653 53 Male <NA> <NA> <NA> <NA> http://~
#> 2 1721 60 Fema~ <NA> <NA> <NA> <NA> http://~
#> 3 1725 57 Male <NA> <NA> <NA> <NA> http://~
#> 4 4 25 Male <NA> <NA> <NA> <NA> http://~
#> 5 512 34 Fema~ <NA> <NA> <NA> <NA> http://~
#> 6 2664 74 Fema~ <NA> <NA> <NA> <NA> http://~
#> 7 2665 88 Fema~ <NA> <NA> <NA> <NA> http://~
#> 8 1391 54 Fema~ <NA> <NA> <NA> <NA> http://~
#> 9 1447 45 Fema~ <NA> <NA> <NA> <NA> http://~
#> 10 1452 44 Fema~ <NA> <NA> <NA> <NA> http://~
#> # ... with 317 more rows, and 10 more variables: snomedCode1 <chr>,
#> # snomedCode2 <chr>, snomedCode3 <chr>, snomedCode4 <chr>,
#> # snomedCode5 <chr>, tissueDescription1 <chr>, tissueDescription2 <chr>,
#> # tissueDescription3 <chr>, tissueDescription4 <chr>,
#> # tissueDescription5 <chr>
```





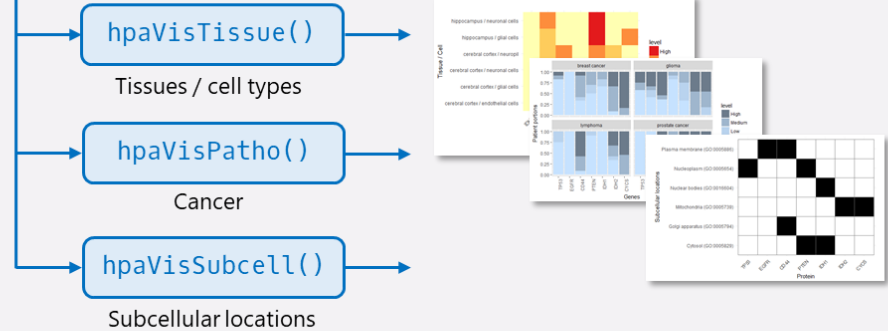
Import, subset and export downloadable datasets



Visualization

hpaVis()

Umbrella function



Individual XML extraction

hpaXml()

Umbrella function

HPA portal individual .xml

hpaXmlGet()

xml nodes

hpaXmlProtClass()

Protein classes

hpaXmlTissueExprSum()

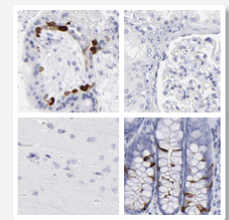
Tissue expression summary
Image URLs

hpaXmlAntibody()

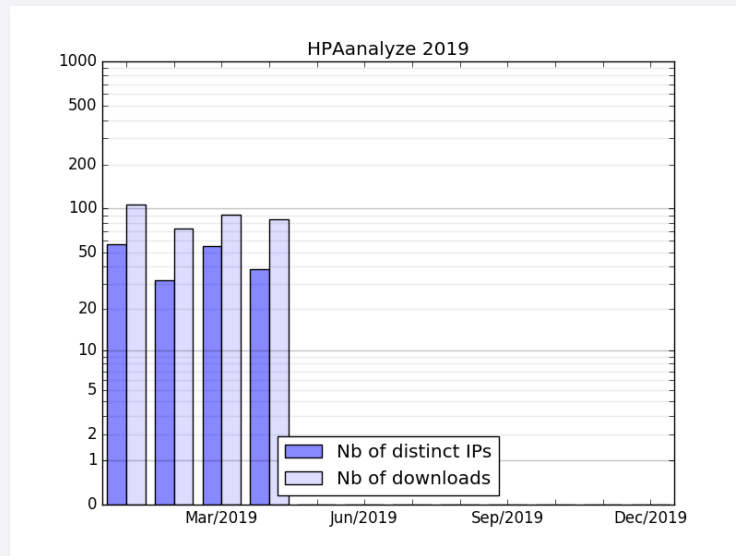
Antibody info

hpaXmlTissueExpr()

Tissue expression details
Image URLs



You're welcome, my 50 users!



It's kinda lonely in the niche...

Issues may go unreported



Hard to get user feedbacks
and no community support



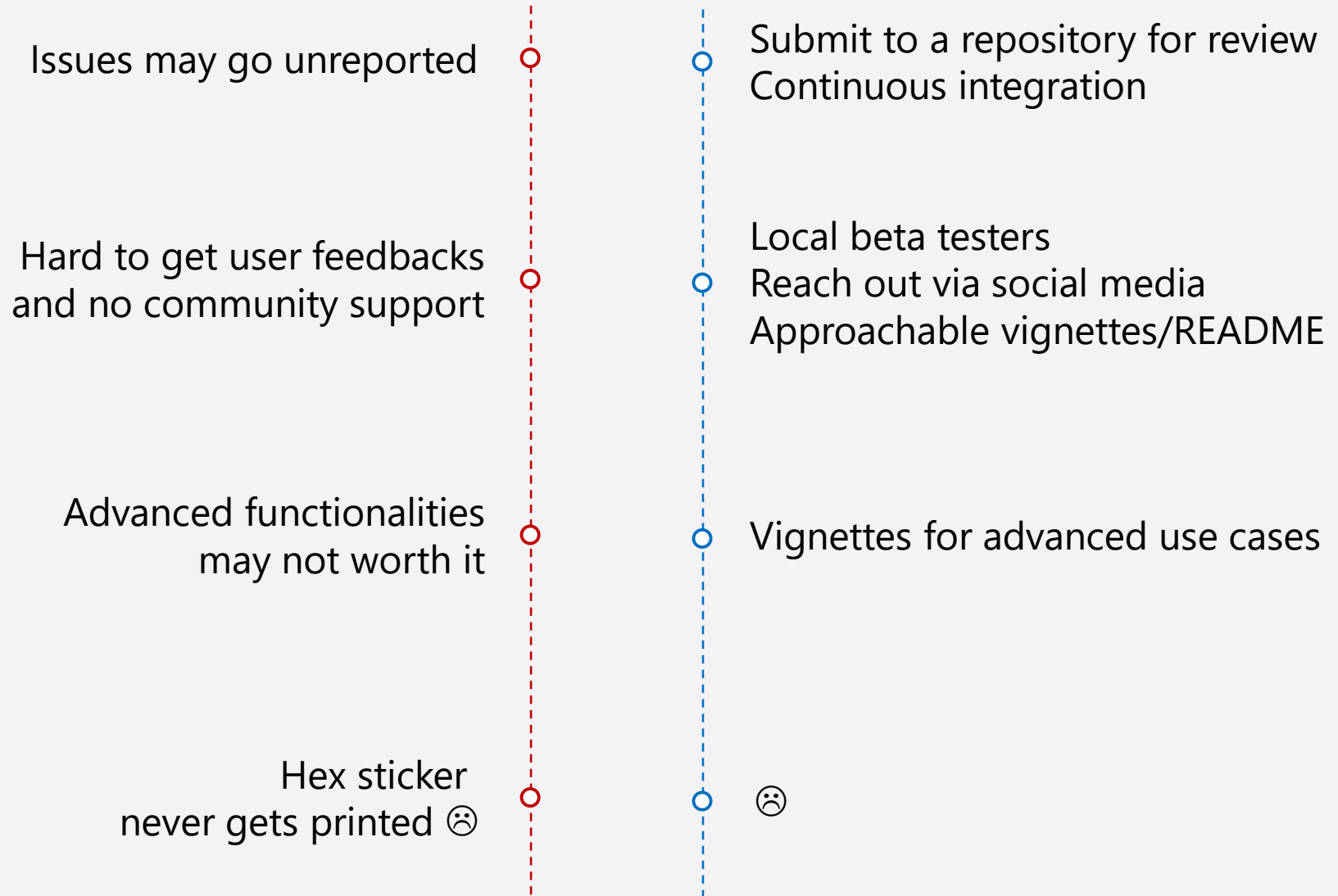
Advanced functionalities
may not worth it



Hex sticker
never gets printed ☹️



It's kinda lonely in the niche...



Acknowledgement



Thank you!

